DOCUMENT RESUME

ED 068 581 TM 002 097

AUTHOR Lehman, Richard S.
TITLE The Use of the Unknown in Teaching Statistics.
PUB DATE 29 Apr 72
NOTE 12p.; Paper presented at the EPA Convention (Boston, Mass., April 29, 1972)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *College Instruction; Computer Programs; Evaluation Methods; *Instructional Technology; Methodology; *Programed Materials; *Statistical Analysis; *Teaching Techniques

ABSTRACT

A technique is described for teaching a second course in statistics at the undergraduate level, but is equally applicable to a first course. The procedure involves the use of individualized data sets, computer prepared to student requests. A package of five programs is presented, which generates the data. They include programs for sampling from bivariate and multivariate normal distribution from independent normal distributions for T-tests and one-way analysis of variance cases, generation of data for rank order correlation, and for analysis of variance designs with two or three independent variables. A separate set of hand-outs to the paper illustrates the use of the program for generating data. (Author/LH)

Presented at EPA Convention, Boston, Mass. 4/29/72

THE USE OF THE UNKNOWN IN TEACHING STATISTICS

Richard S. Lehman
Franklin and Marshall College

I should start by commenting that neither the general idea of using an unknown in teaching statistics nor the use of the word "unknown" itself in this context was original with me. I must acknowledge a debt to James Coleman and Doris Entwistle of Johns Hopkins University for I guess originating the concept, and to Sally Sedelow of the University of Kansas for initially planting the idea in my head.

What I want to talk about is a way of teaching statistics. There are some cautions which I should make right at the outset about this particular way that I have of running my course. The first caution is that it is terribly time-consuming for the instructor. The second warning is that a course using individualized data sets (which is what this is really all about) makes the course entirely computer-dependent. You must have access to a highly reliable computer system, and hopefully at least one highly reliable teaching assistant.

On the positive side, the structure of a statistics course, as I want to talk about it, has several benefits which I feel outweigh the costs. One of these is the preparation of students to do statistics in the context of real experiments in a much more thorough and meaningful way. The students come away from the course with a much better understanding of how to design experiments, and how to analyze and interpret data, than they would in any other way of teaching statistics that I have seen or used. In addition,

ABSTRACT

THE USE OF THE UNKNOWN IN TEACHING STATISTICS

This paper describes a technique used in teaching a second course in statistics at the undergraduate level. ~~Some~~ The techniques discussed here are equally applicable to a first course. The procedure involves the use of individualized data sets, computer prepared to student requests. A package of five programs is presented, which generates the data. They include programs for sampling from bivariate and multivariate normal distribution from independent normal distributions for T-tests and one-way analysis of variance cases, generation of data for rank order correlation, and for analysis of variance designs with two or three independent variables. A separate set of hand-outs to the paper illustrates the use of the program for generating data.

for the first time in 8 years of teaching statistics, I find the students are getting excited about statistics--turned on by statistics per se rather than viewing it as a necessary, but hard to grasp and use, tool.

The general concept that I am working with is a highly individualized approach. The class itself meets as a normal class with lectures, tests, assigned readings, etc. Instead of completing assigned problem sets during the course of the semester, however, the student is required to turn in a certain number of what I call experimental designs. I'll comment more about what a design consists of in a few moments.

As a part of the design the student requests a set of data. Through the use of a set of data generator programs, it is possible for me to supply to each student data specific to his experimental design. Variables are labeled, subjects may be named in some instances, maximum and minimum values of variables may be specified, etc.

When the student has received his data and his graded experimental design, he then proceeds to conduct the data analysis as if it were data from an actual experiment, and then hands in the data analysis, including not only the computations but also a short discussion interpreting what he has found.

I have found that the students become quite involved with the hypothetical experiments that they turn in, and are very disappointed when the data do not work out as expected, and very elated when they do.

There is another benefit here. It is one of the few "experimental" situations in which the true population parameters are known (--I know them, the students do not). It is therefore possible for me, in grading the data analyses, to tell the student if in fact he has made a Type I or Type II error. The students have found this very interesting, and it has certainly helped to clarify one of the more difficult concepts for undergraduate students to grasp.

To put the course in a somewhat broader context, I should describe how it fits into the psychology curriculum at Franklin and Marshall. The F&M psychology program is a highly research-oriented, basic psychology program. All majors are required to complete two semesters of laboratory introductory psychology, a two-semester statistics and methodology sequence, four intermediate survey courses, and four advanced-level laboratory courses which involve conducting individually designed research projects in various areas. In addition, seniors also complete either an independent research project or a semester-long senior seminar.

The statistics course I am talking about is the second semester of the two-semester sequence. Students come into my course with a considerable background in the nature of experimental research in psychology, with an understanding of elementary descriptive statistics, an introduction to hypotheses testing, and statistical tools through a two-group t-test.

The second-semester course begins with a review of hypothesis testing, including a strong emphasis on the concept of power and Type I and Type II errors. We proceed then through one- and two-way analysis of variance, correlation-regression and multiple correlation-regression. Students are introduced to nonparametric analogs to the parametric procedures as appropriate throughout the course.

Mechanics

I have taught the course for two semesters using the unknown approach. I am still experimenting with what is the best structure in terms of requirements. This semester, students are required to complete a minimum of six experimental designs, one each in the areas of two group designs, one-way analysis of variance, non-parametric, one-way analysis of variance, correlation-regression, and two experimental designs in higher level analysis of variance involving two or more independent variables. They have the option of completing as many as two designs in each of the first four areas, and four in the latter.

An experimental design consists of a one-page (hopefully) description of an experimental situation identifying the independent and dependent variables, the nature and number of subjects, the control of extraneous variables, etc. With the experimental design, the student turns in one of three possible data request forms. A data analysis must contain computations and an interpretation of the results in terms of the original design.

Programs

Let me now outline briefly the set of five programs which I have written to generate data for students' experimental designs. All of the programs are at the moment written in the BASIC language and operate interactively through a teletype to an RCA Spectra 70/46. Copies of the programs are available to anyone who is interested. I will be pleased to send listings of the programs, or if you happen to have a 70/46, I will supply the programs on cards.

During the coming summer, I hope to be able to convert all of the programs from BASIC into FORTRAN so that they will not only be more readily transferable to other installations, but will also be able to run in a batch rather than an interactive mode.

The programs are all organized in essentially the same way. Each program requests some identification for a particular set of data. Then it requests, for each group or for each independent variable, the parameters of the population or populations to be sampled. Usually the requests are in the form of expected values and variances. In addition, there is an option of specifying the range for the data values, and whether or not the data values may have decimal parts.

The output of all of the programs is organized similarly. There is first a set of output for the student. It contains the ID of the data set, the names of the different groups, variables, or levels, and the actual data values.

The second part of the output is an instructor's record subdivided into two categories--parameters and statistics. In the parameters section, the program prints the parameter values that were input, describing the populations sampled. The statistics section lists summary values for the actual data which were given to the student, usually with the appropriate statistical test applied. For example, in the single independent variable data generator, the sum of squares, mean, and variance is given for each group. For a two-group design, the computations for the two-group t-test follow. If there are more than two groups, the summary table for analysis of the variance is printed.

The different programs are:

DATARANK: This program generates data for a rank order correlation.

DATAGEN 1: This is the single independent variable data generator. From 1 to 10 independent groups of data, each with varying group sizes, may be generated.

CORRDATA: This is a two variable data generator for bivariate normal sampling, or for two correlated groups.

DATAGEN3: This program generates 2 and 3 way analysis of variance data. Repeated measures, random and fixed factors, nested variables, etc. are permitted.

MULTDATA: This program (which I admit is still in the writing) samples from a multivariate normal population of up to 5 variables. It is used for generating data for multiple correlation and regression, and also for 3 or more groups when the groups are correlated. Input includes expected values and variance/covariance matrix.

Summary

I have found that this approach to statistics is far superior to any other method that I've tried. The students like it, and learn a great deal. In the other courses that they take after this, they are much better able to deal with the problems of setting up research designs, and calculating and interpreting data than they have been in the past.

The technique of the "unknown" has been applied to other statistics courses whose content is not as advanced as mine. The data generator programs can be used in other ways as well; for example, to provide data for individual students doing more standard statistics problem sets. The programs could be run by the students themselves to generate their own data, and have a way of checking their computation. It would be interesting to use the programs in a statistics laboratory setting, where various population characteristics could be explored quickly to illustrate their effects on the obtained samples.

As I said before, I will be pleased to supply the programs and any necessary additional information to anyone who is interested. In return, I would like to learn of the approaches that others might take in using the programs, and the success that may have.

Thank you very much.

THE USE OF THE UNKNOWN
IN TEACHING STATISTICS

Richard S. Lehman

Franklin & Marshall College
Lancaster, Pa. 17602

Handout for EPA presentation, April 29, 1972

Example Data Request

DATA REQUEST FORM SINGLE I.V. STUDIES

Complete all blanks except those enclosed in dotted lines.

Name John Q Student

I.D. John Q Student II-2

Number of groups 3

Are groups ✓ independent or ___ correlated?

Complete the following for each group.

GROUP 1

No. observations in this group 10

Group I.D. (15 characters) High SES

Range of values: 70 to 150

Normal distribution? ✓ yes ___ no
If no, the long tail is to the
___ left ___ right

Can data have decimals? ___ yes ✓ no

μ 110
σ 15
min 70
max 150

GROUP 2

No. observations in this group 12

Group I.D. (15 characters) Middle SES

Range of values: 60 to 150

Normal distribution? ✓ yes ___ no
If no, the long tail is to the
___ left ___ right

Can data have decimals? ___ yes ✓ no

μ 90
σ 16
min 60
max 150

GROUPS 3

No. observations in this group 10

Group I.D. (15 characters) Low SES

Range of values: 60 to 150

Normal distribution? ✓ yes ___ no
If no, the long tail is to the

___ left ___ right

Can data have decimals? ___ yes ✓ no

μ 90
σ 16
min 60
max 150

If you have more than three groups, please use another form, completing only the number of group description blanks required for your study. Attach the sheets securely to your experimental design.

Sample Run of DATAGEN1

```
SINGLE I.V.DATA GENERATOR,INDEPENDENT GROUPS   2/14/72
INSTRUCTIONS (Y,N)?N
INPUT BY DATA STATEMENTS (Y,N)?N
DATA SET ID?JOHN Q STUDENT  11-2
NUMBER OF GROUPS?3
ENTER N,GROUP ID,E(X), SD(X),MIN(X), MAX(X) FOR GROUP 1
?10,HIGH SES,110,15,70,150
ENTER N,GROUP 1D,E(X),SD(X),MIN(X),MAX(X) FOR GROUP 2
?12,MIDDLE SES,90,16,60,150
ENTER N,GROUP 1D,E(X), SD(X),MIN(X), MAX(X) FOR GROUP 3
?10,LOW SES,90,16,60,150
DECIMALS (Y,N)?N
DATA SET 1D-----JOHN Q STUDENT  1-2
GROUP 1 (HIGH SES) N = 10
 110            134          102        100           104
 84             126          113        77            125
GROUP 2 (MIDDLE SES) N = 12
 119            93           80         75            113
 103            69           108        103           83
GROUP 3 (LOW SES) N = 10
 71             91           69         77            91
 93             98           95         82            95
INSTRUCTORS RECORD
DATA SET ID  JOHN Q STUDENT  11-2
PARAMETERS-----
GROUP 1 (HIGH SES)
  E(X) = 110  SD(X) = 15    RANGE 70 - 150
GROUP 2 (MIDDLE SES)
  E(X) = 90  SD(X) = 16     RANGE 60 - 150
GROUP 3 (LOW SES}
  E(X) = 90  SD(X) = 16     RANGE 60 - 150
STATISTICS-----
GROUP 1 (HIGH SES) N = 10
SUM = 1075    MEAN = 107.5
SUM X**2 = 118531       VAR = 296.848 S.D. = 17.2293
GROUP 2 (MIDDLE SES) N = 12
SUM = 1118    MEAN = 93.1667
SUM X**2 = 107056       VAR = 241.309 S.D. = 15.5341
GROUP 3 (LOW SES) N = 10
SUM = 862     MEAN = 86.2
SUM X**2 = 75320        VAR = 101.563 S.D. = 10.0778

       SUMMARY TABLE
SOURCE         SS            DF         MS            F
BETWEEN         2370.19       2          1185.09       4.99544
WITHIN          6879.81       29         237.235
    TOTAL       9250          31
```

Input

Student data

Instructors Record

Data Generator Programs

DATARANK: This program generates data for a rank order correlation.

DATAGEN1: This is the single independent variable data generator. From 1 to 10 independent groups of data, each with varying group sizes, may be generated.

CORRDATA: This is a two variable data generator for bivariate normal sampling, or for two correlated groups.

DATAGEN3: This program generates 2 and 3 way analysis of variance data. Repeated measures, random and fixed factors, nested variables, etc. are permitted.

MULTDATA: This program samples from a multivariate normal population of up to 5 variables. It is used for generating data for multiple correlation and regression, and also for 3 or more groups when the groups are correlated. Input includes expected values and variance/covariance matrix.